

Initiatives in 100 GE for Fermilab R&D

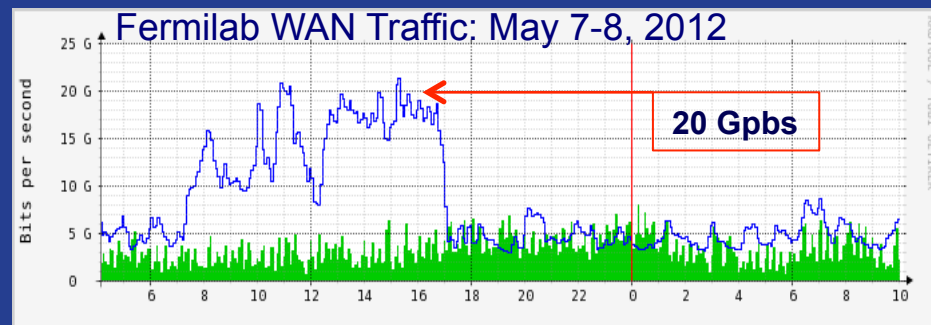
Parag Mhashilkar & Gabriele Garzoglio
Grid and Cloud Computing Department
Computing Sector, Fermilab

Overview

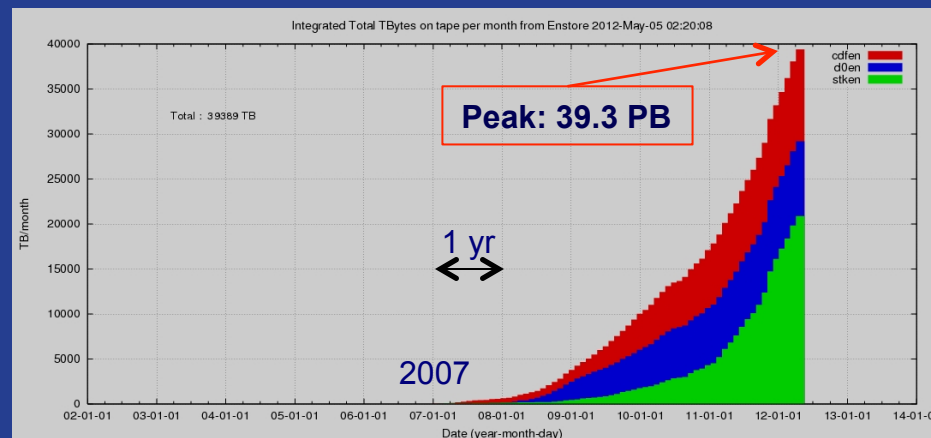
- Fermilab's interest in 100 GE
- Results from the ANI testbed
- Future program
- 100 GE infrastructure at Fermilab

Fermilab Users and 100 GE

- Decades long dependence on sustained, high speed, large and wide-scale distribution of and access to data
 - High Energy Physics community
 - Multi-disciplinary communities using grids (OSG, XSEDE)
- Figures of merit
 - 40 Petabytes on tape, today mostly coming from offsite
 - 140Gbps LAN traffic from archive to local processing farms
 - LHC peak WAN usage at 20-30 Gbps



Compact Muon Solenoid (CMS) routinely peaks at 20-30 Gbps.



Goals of 100 GE Program @Fermilab

- End-to-end experiment analysis systems include a deep stack of software layers and services.
- **Need to ensure these are functional and effective at the 100 GE scale.**
 - Determine and tune the configuration to ensure full throughput in and across each layer/service.
 - Measure and determine efficiency of the end-to-end solutions.
 - Monitor, identify and mitigate error conditions.

High Throughput Data Program (HTDP) at Fermilab

- **Mission:** prepare the Computing Sector and its stakeholders for the 100GE infrastructure and put Fermilab in a strategic position of leadership.
- Establish collaborations with stakeholders, computing facilities, scientific communities, and institutions, to coordinate a synergistic program of work on 100GE.
- The program includes technological investigations, prototype development, and the participation to funding agency solicitations.
- The ANI has been the major testbed used since last year in close partnership with ESN

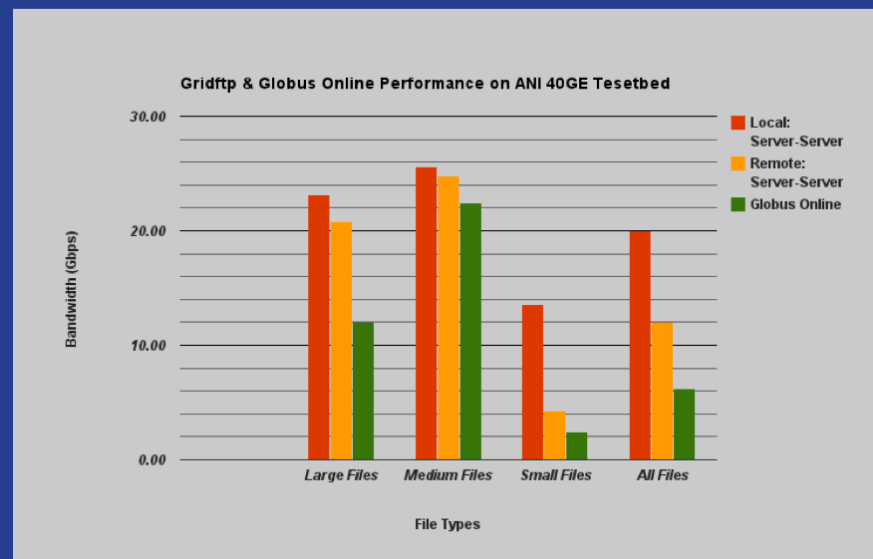
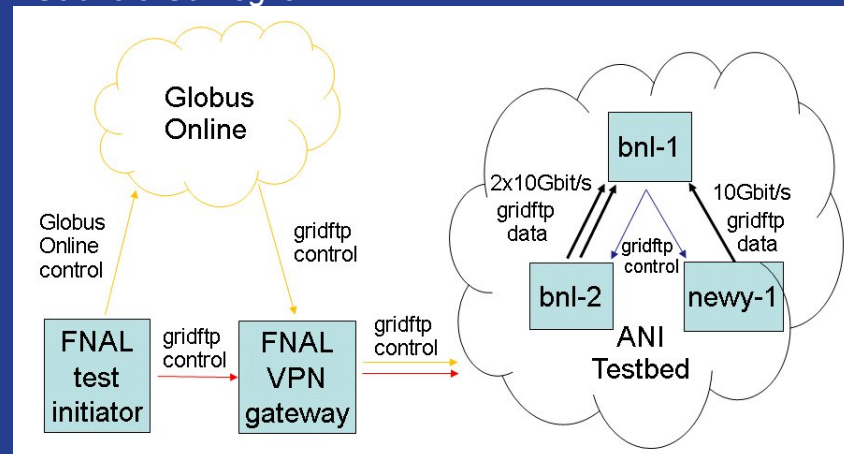
Ongoing Program of Work

- 2011: ANI Long Island MAN (LIMAN) testbed.
 - Tested GridFTP and Globus Online for the data movement use cases of HEP over 3x10GE.
- 2011-2012: Super Computing 2011.
 - Demonstration of fast access to ~30TB of CMS data from NERSC to ANL using GridFTP.
 - Achieved 70 Gbps
- Currently: ANI 100GE testbed.
 - Tuning parameters of middleware for data movement: xrootd, GridFTP and Globus Online.
 - Achieved ~97Gbps
- Summer 2012: 100GE Endpoint at Fermilab
 - Plan to repeat and extend tests using CMS current datasets.

Experience on the ANI LIMAN Testbed

Work by Dave Dykstra w/ contrib. by Raman Verma & Gabriele Garzoglio

- Testing with GridFTP using 3x10GE in preparation for 100GE on ANI Testbed.
- Characteristics:
 - 300GB of data split into 42,432 files (8KB – 8GB; varied sizes).
 - Aggregated 3 x 10Gbit/s link to Long Island test end-point.
- Results:
 - Almost equal throughput for Globus Online (yellow) as for direct GridFTP (red) for medium-size files.
 - Increased throughput by 30% through increasing concurrency and pipelining on small files.
 - Auto-tuning in Globus Online works better for medium sized files than for large files.

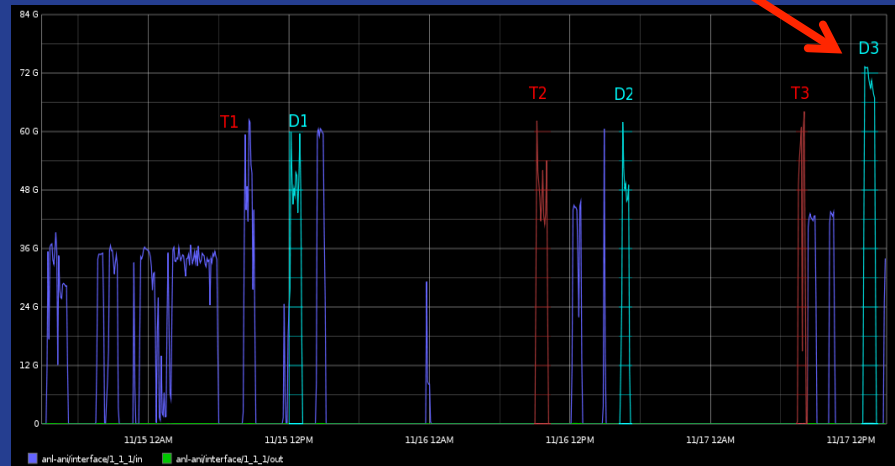
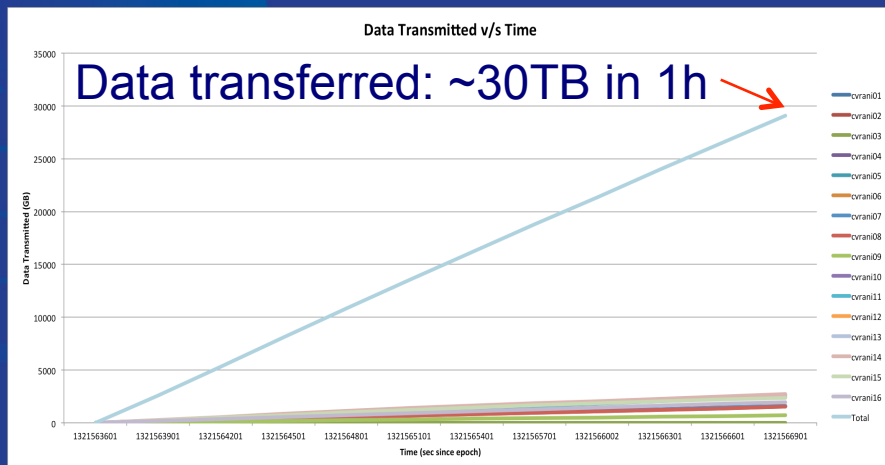


Super Computing 2011

- Test transfer of CMS experiment data between NERSC and ANL over 100 GE network.
- Characteristics:
 - 15 server / 28 client nodes (multi-cores, 48 GB RAM, 10Gbps)
 - 2 globus-url-copy (GUC) clients / server

Work by Parag Mhashilkar, Gabriele Garzoglio (Fermilab) and Haifeng Pi (UCSD)

	GUC/ core	GUC streams	GUC TCP Window Size	Files/ GUC	MAX BW	Sustain BW
T1	-	-	-	-	-	-
D1	1	2	Default	60	65	50
T2	1	2	2MB	1	65	52
D2	1	2	2MB	1	65	52
T3	4	2	2MB	1	73	70
D3	4	2	2MB	1	75	70

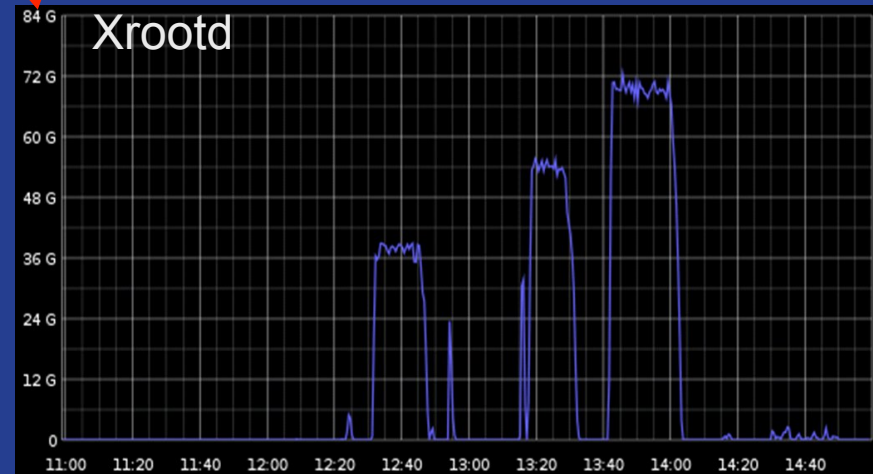
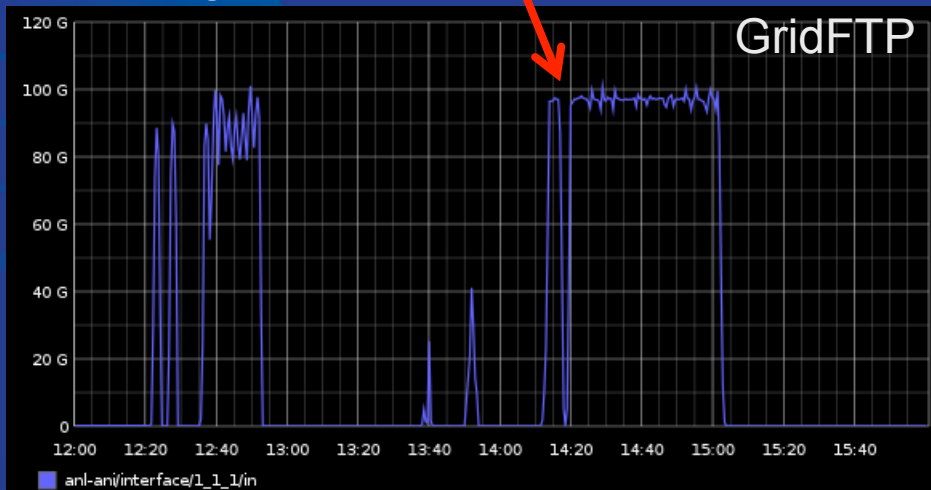


Using the ANI 100G Testbed

- Data Movement over Xrootd, testing LHC experiment (CMS / Atlas) analysis use cases.
 - Clients at NERSC / Servers at ANL
 - Using RAMDisk as storage area on the server side
 - Challenges
 - Tests limited by the size of RAMDisk
 - Little control over xrootd client / server tuning parameters
- Achieved 97 Gbps with limited testing of GridFTP

Work by Hyunwoo Kim (Fermilab)

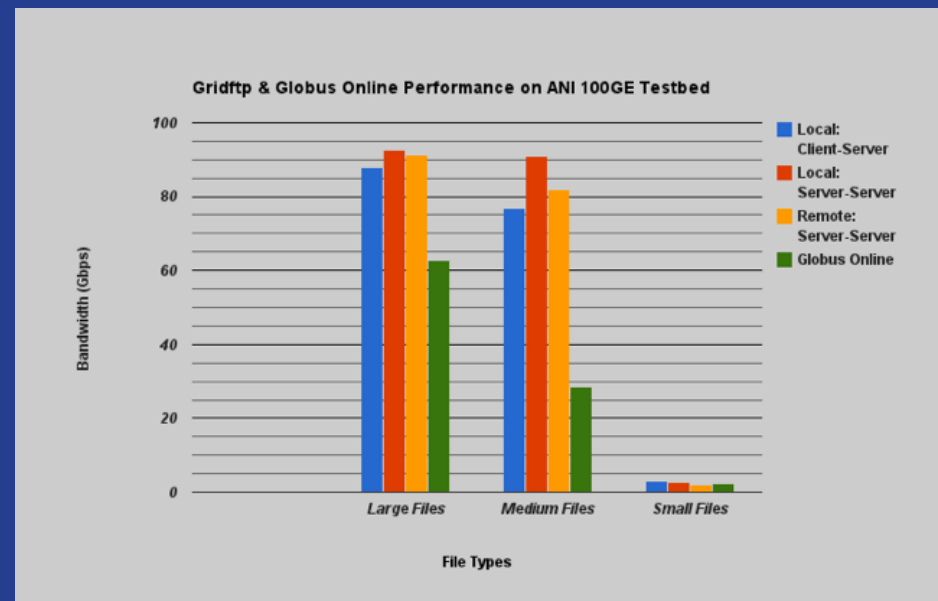
# Clients	Input File 512 MB	Input File 1 GB	Input File 2 GB	Input File 4 GB
1	~12 Gbps	~18 Gbps	~26 Gbps	~32 Gbps
2	~22 Gbps	~37 Gbps	~40 Gbps	~56 Gbps
4	~42 Gbps	~56 Gbps	~73 Gbps	~77 Gbps
8	~60 Gbps	~75 Gbps	~80 Gbps	-



Gridftp & Globus Online Performance on ANI 100GE Testbed

- Data Movement using gridftp.
 - Clients at NERSC / Servers at ANL
 - Dataset split into 3 sets, size increasing in powers of 2: Small(8KB - 4MB), Medium(8MB -1G), Large(2, 4, 8 GB)
- Large files transfer performance ~ 92Gbps
- Small files transfer performance - abysmally low
- Issues uncovered on 100G Testbed
 - Small is relative
 - Medium is new small

	Local: Client-Server	Local: Server-Server	Remote: Server-Server	Globus Online
Large	87.92 Gbps	92.74 Gbps	91.19 Gbps	62.90 Gbps
Medium	76.90 Gbps	90.94 Gbps	81.79 Gbps	28.49 Gbps
Small	2.99 Gbps	2.57 Gbps	2.11 Gbps	2.36 Gbps

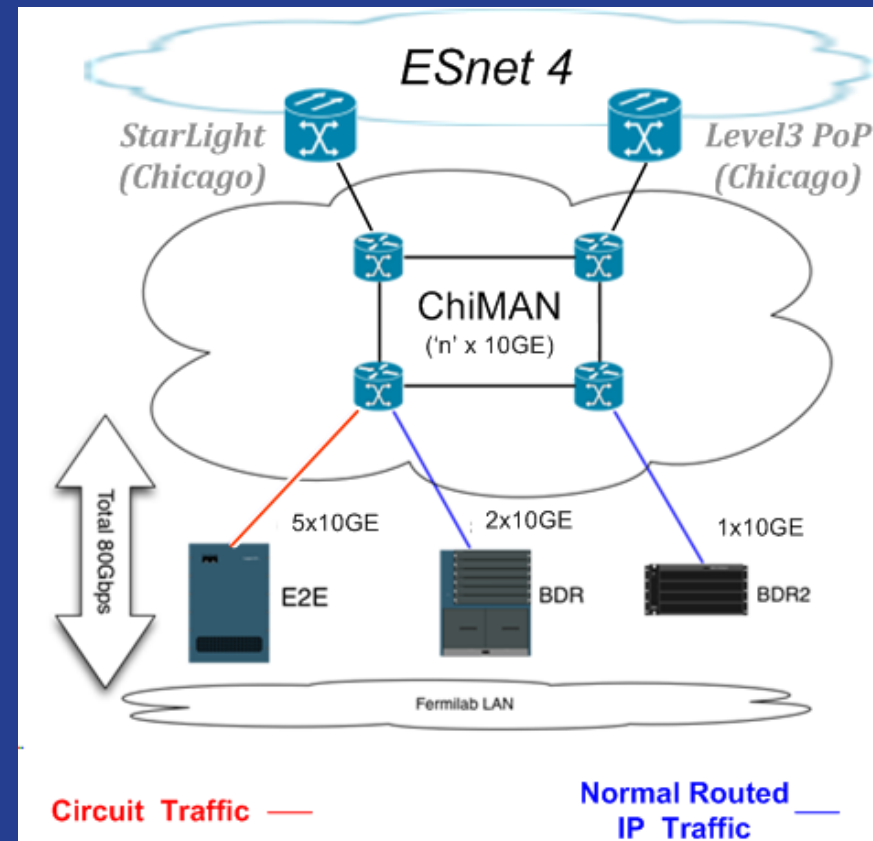


Current Plans & Constraints

- ANI 100G testbed
 - Current time window: until Aug 2012
 - Complete tests of Xrootd, GridFTP, and Globus Online
 - Test Squid for condition data access
- Without an ANI extension, we'll delay or cancel the testing of technologies used by other Fermilab stakeholders:
 - Luster, IRODS, CVMFS, dCache.
 - This would mean an increased risk to the stakeholders.
- 100GE production endpoint coming to Fermilab (see next slides)
 - Expecting 100 GE capabilities in summer 2012.
 - Creating a local testbed connecting to ANI.
 - Continue testing of middleware technologies defined by stakeholders.

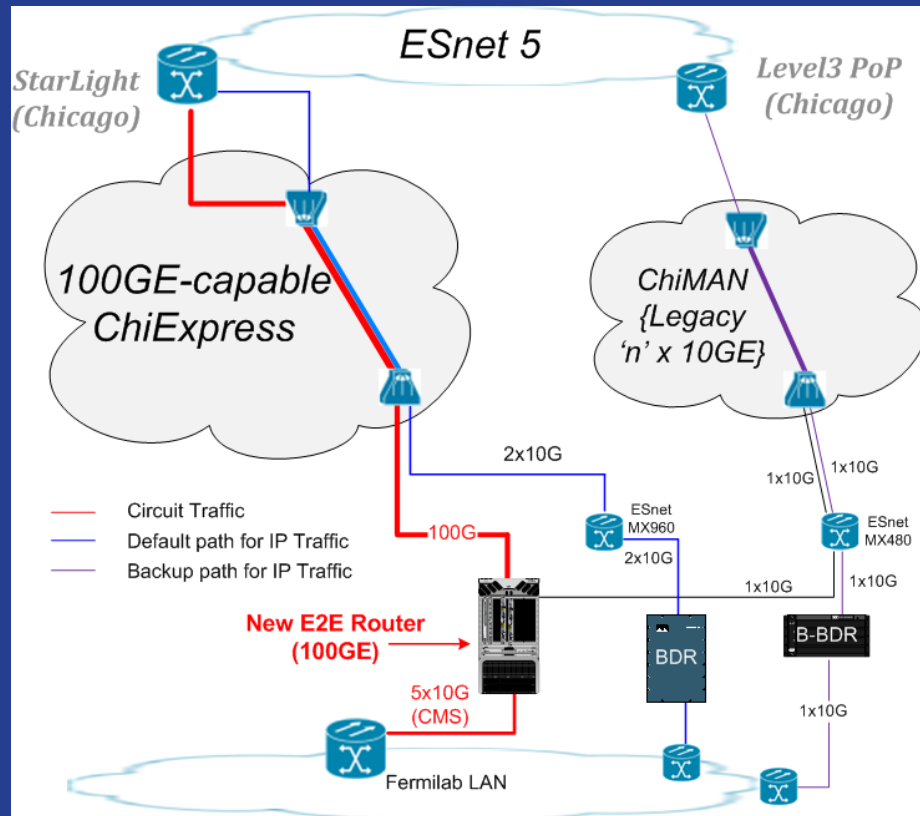
Current Fermilab WAN Capabilities

- Metropolitan Area Network provides 10GE channels:
 - Currently 8 deployed
- Five channels used for circuit traffic
 - Supports CMS WAN traffic
- Two used for normal routed IP traffic
 - Backup 10GE for redundancy
 - Circuits fail over to routed IP paths



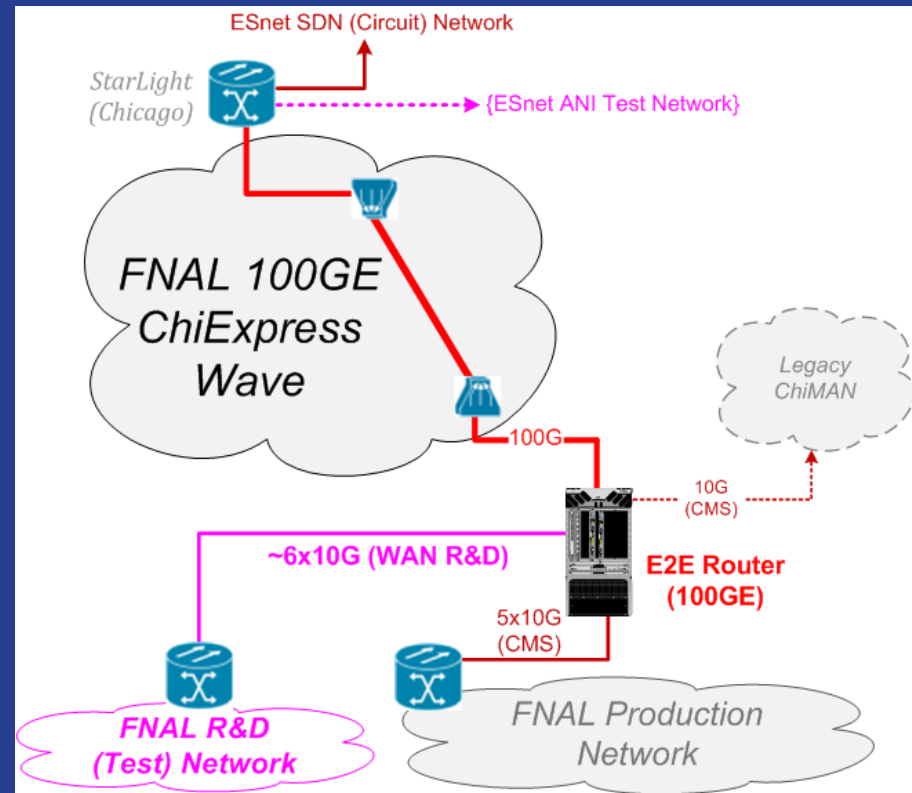
100GE WAN capability is coming

- ESnet deploying 100GE MAN as part of ESnet5
 - One 100GE wave for FNAL
 - Also 2x10GE channels for routed IP traffic
- 100GE wave will be used to support circuit traffic
- Legacy 10GE MAN will remain for diversity
 - Backup routed IP path
 - One 10GE circuit path, too



Use of 100GE Wave for FNAL R&D

- 100GE wave will support 4x10GE circuit paths
 - Excess capacity available for WAN R&D activities
- Planning ~6 x 10GE link to FNAL R&D network
 - Network will host 10GE test/development systems
 - Possibly 40GE systems later
- Anticipate WAN circuit into ESnet ANI test bed



Summary

- Fermilab has a program of work to test 100GE network for its scientific stakeholders
- The collaboration with ANI and ESNet has been central to this program
- The current timeline for ANI is not sufficient to evaluate all technologies of interest to the Fermilab stakeholders
- Fermilab will have 100GE capability in the summer 2012 – planning for involvement with ANI